Technical white paper

HP Hardware Accelerated Graphics for Desktop Virtualization

Technology and implementation overview on ProLiant servers

Table of contents

Purpose of this document 2
Abbreviations and naming conventions 2
Introduction to HP Enterprise Client Virtualization Reference Architecture 3
The HP WS460c Gen8 Graphics Server Blade 3
The HP Multi-GPU Carrier Card 4
Hardware accelerated graphics for desktop virtualization concepts and technology 4
Bare Metal OS 4
Pass-through GPU 5
Software Virtualized GPU 6
Hardware Virtualized GPU—True virtual GPU 7
Graphics accelerated desktop sessions and application virtualization 8
Planning consideration for implementing hardware accelerated desktop virtualization technologies 9
Determining the right GPU and platform for your use case 9
Graphics accelerated desktop virtualization platform feature comparison 10
Calculating VM/user density 11
HP WS460c Gen8 Graphics Server Blade configuration and support guidelines 13
HP WS460c Gen8 Graphics Server Blade documentation 13
Understanding WS460c Gen8 Graphics Server Blade Core configurations and options 13
WS460c Graphics Server Blade OS, hardware, and firmware guidelines and support 14
Required WS460c BIOS configurations 18
Platform specific version and configuration recommendations 21
Citrix XenServer Direct Map system configuration recommendations 21
VMware system configuration recommendations 22
Microsoft Hyper-V/RemoteFX system configuration recommendations 23
Appendix A— Alternative remoting protocols and brokers 26
Appendix B—Choosing the best graphics hardware and client virtualization solution 27
Resources 30

Click here to verify the latest version of this document

Purpose of this document

- Provides an overview of the technical concepts and practical implementation of graphics hardware accelerated virtual desktop technologies.
- Gives IT decision makers, architects, and implementation specialists an overview of how HP and its virtualization partners approach and implement hardware accelerated graphics for virtual desktop.

Abbreviations and naming conventions

Table 1. Abbreviations and terminology used in this document

Convention	Definition
Broker/Connection Broker	Manages connections between end user devices and remote systems/desktop
Bare Metal OS	Operating system installed directly on system, not virtualized
CV	Client Virtualization
DAS	Direct Attached Storage
GPU	Graphical Processing Unit (Graphics Card). It is important to note that some graphics cards have more than one GPU processor
GPU Compute	Synonymous to GPGPU. See GPGPU
GPGPU	General-Purpose Graphics Processing Unit. GPU technology that performs application computation traditionally handled by the CPU
HDX	Citrix set of advanced desktop remoting technologies to deliver a High Definition Experience
HDX 3D Pro	Feature of XenDesktop for delivering high-end 3D professional graphics
Hypervisor	Virtualization host platform (VMware ESXi, Microsoft® Hyper-V, Citrix XenServer)
ICA	Independent Computing Architecture protocol (Part of Citrix HDX technologies)
OA	Onboard Administrator
Pass-Through Graphics	Technology that allows hypervisor to directly pass-through a graphics card device to a VM (for example a GPU)
PCoIP	Teradici remote desktop protocol used in VMware View
RBSU	ROM Based Setup Utility (HP Server BIOS)
RDS	Microsoft Server feature that enables users to connect to virtual desktops, session-based Desktops and RemoteApp programs
RDP	Microsoft Remote Desktop Protocol
RemoteFX	Microsoft set of advanced desktop remoting technologies
RFX	Microsoft RemoteFX
RGS	HP Remote Graphics Software
SAN	Storage Area Network
softGPU	Software emulated GPU
VDI	Virtual Desktop Infrastructure
vGPU	vGPU has a different definition depending on platform used: • Microsoft RemoteFX vGPU: Microsoft implementation of software virtualized GPU • VMware vGPU: VMware implementation for software virtualized GPU • NVIDIA vGPU: Hardware virtualization of the GPU, NVIDIA GRID vGPU
vSGA	VMware specific terminology Software Virtualized GPU (API Capture Model)
vDGA	VMware specific terminology for GPU pass-through
VM	Virtual Machine

Introduction to HP Enterprise Client Virtualization Reference Architecture

HP Client Virtualization (CV) Gen8 Enterprise Reference Architectures have been tested and optimized with select combinations of storage, networking and server blades, and are proven to efficiently handle today's CV workloads. In addition to capitalizing on the increased performance and user density of the HP ProLiant WS460c Gen8 Graphics Server Blade, the latest solutions also benefit from HP's new Proactive Service and Support. HP Client Virtualization can help customers achieve the goals of IT and workforce support, without compromising performance, operating costs, information security, and user experience with reference architectures.

Simplicity: With an integrated data center solution for rapid installation and startup and easy ongoing operations

Optimization: A tested solution with the right combination of compute, storage, networking, and system management tuned for Client Virtualization efficiency

Flexibility: With options to scale out to meet precise customer requirements

Server performance: With options that use hardware accelerated graphics to meet the demands of customers high end user needs

VDI is one possible implementation of the CV reference architecture. HP Graphics Server Blades and server based-computing also fit the CV model.

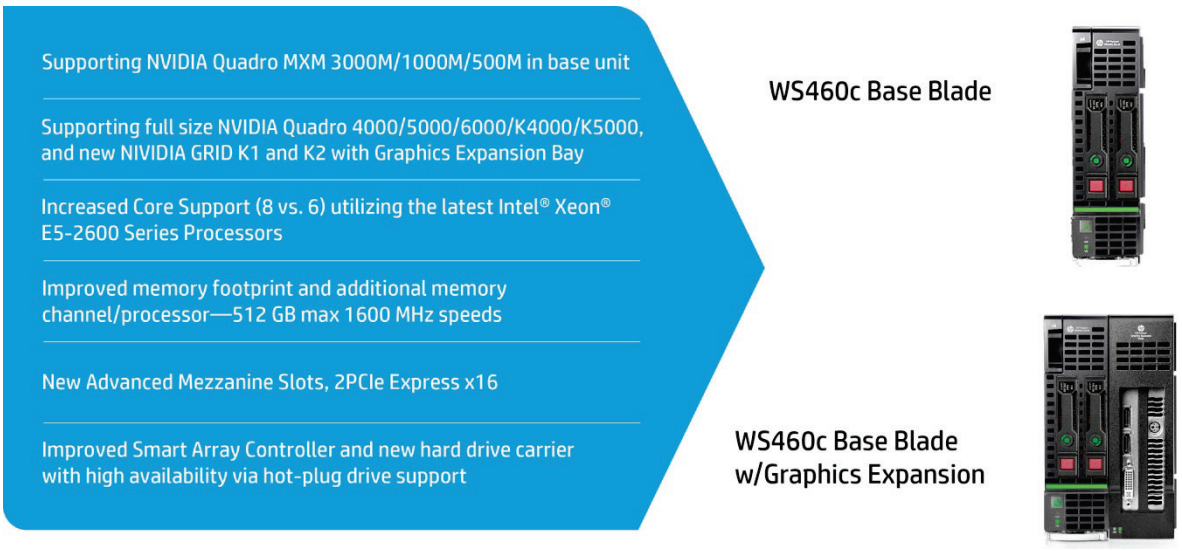
For more information on HP CV Gen8 Enterprise Reference Architectures, visit hp.com/go/cv.

The HP WS460c Gen8 Graphics Server Blade

The HP ProLiant WS460c Gen8 Graphics Server Blade (Formerly WS460c Gen8 Workstation Blade) has been at the cutting edge of workstation computing for years by allowing you to centralize your organization's workstations in the data center. Rather than placing the workstation's computing power at the user's desk, the computing power—in the form of a server blade—is moved to the data center, where servers can be more easily, securely, and economically managed. The results are improved uptime and business continuity, enhanced data center security, and reduced IT costs. The users that would benefit from such implementation might include digital content and web creation, financial traders, Oil and Gas engineering or any heavy compute users. The WS460c Graphics Server Blade provides a local workstation experience to end users over the network using the one of the common remoting protocols such as Citrix HDX 3D, VMware PCoIP, and Microsoft RemoteFX protocols. Traditionally the Graphics Server Blade has been a Bare Metal 1:1 solution, meaning that a client operating system was loaded on blade for a single user 1:1 environment.

The WS460c Gen8 Graphics Server Blade breaks the mold again with industry first technology supporting up to eight GPU's per blade, enhanced memory footprint and speeds, full PCIe x16 GPU support, all on the proven HP ProLiant Gen8 architecture. The WS460c is ideal for bare metal or virtualized multi-tenancy high-end graphics users. All these features enable users to complete large model visualizations with uncompromised workstation-class performance as well as provide media-rich graphics to PC users.

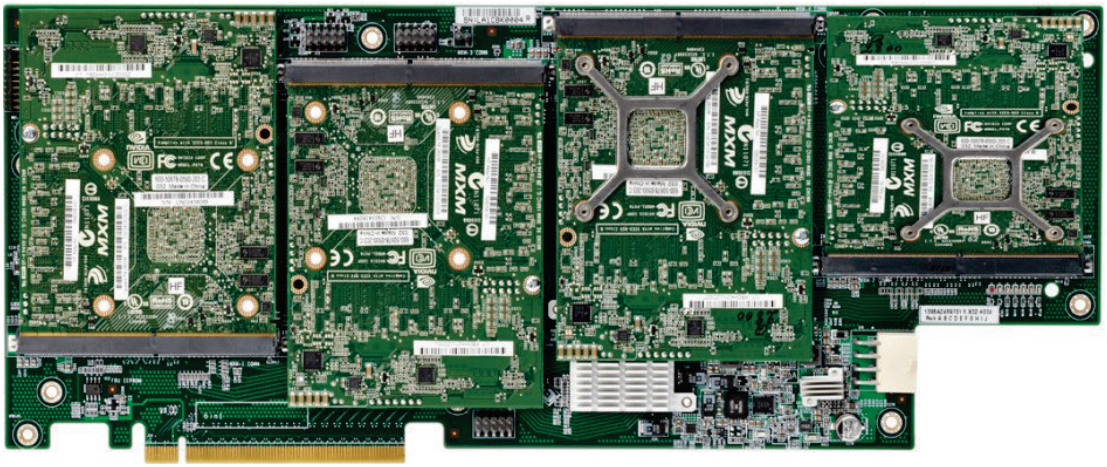
Figure 1. HP ProLiant WS460c Gen8


Supporting NVIDIA Quadro MXM 3000M/1000M/500M in base unit
Supporting full size NVIDIA Quadro 4000/5000/6000/K4000/K5000, and new NIVIDIA GRID K1 and K2 with Graphics Expansion Bay
Increased Core Support (8 vs. 6) utilizing the latest Intel® Xeon® E5-2600 Series Processors
Improved memory footprint and additional memory channel/processor—512 GB max 1600 MHz speeds
New Advanced Mezzanine Slots, 2PCIe Express x16
Improved Smart Array Controller and new hard drive carrier with high availability via hot-plug drive support
WS460c Base Blade
WS460c Base Blade w/Graphics Expansion

The HP Multi-GPU Carrier Card

HP Multi-GPU Carrier card (see figure 2) for the WS460c Gen8 Graphics Server Blade is an industry first MXM (small form factor GPU card) carrier card technology with four MXM slots making the HP Graphics Server Blade the highest density GPU blade platform on the market. Providing high-density, high-end 3D graphics for GPU accelerated desktop virtualization with support for up to eight GPU's (256 GPU's per rack) in a blade form factor to support all accelerated graphics for desktop virtualization technologies.

Figure 2. HP Multi-GPU Carrier Card[1]



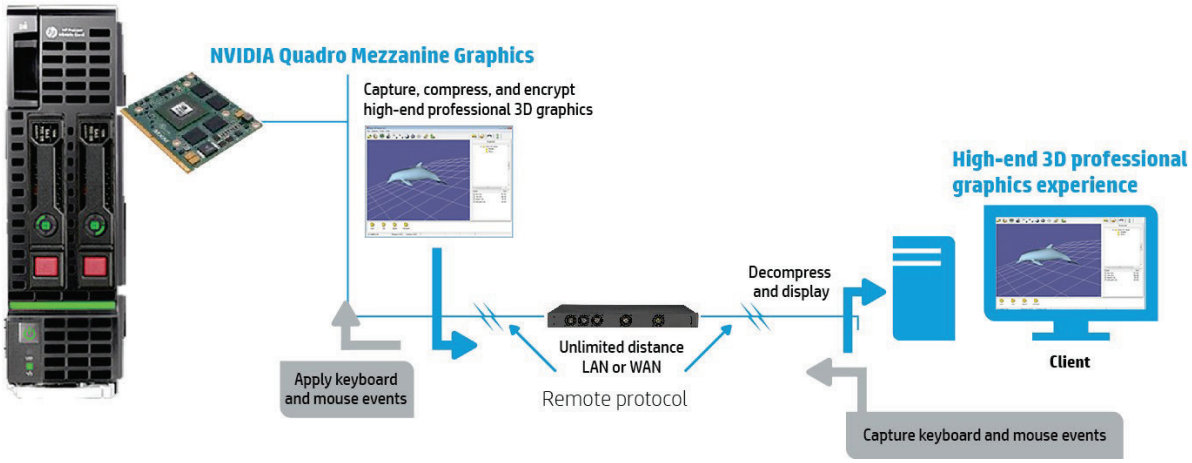
Hardware accelerated graphics for desktop virtualization concepts and technology

In this section, we will provide a conceptual overview of the technologies behind hardware accelerated graphics for desktop virtualization. We will at a high level discuss the differences between the technologies as well as how the major desktop virtualization providers implement these technologies into their products.

Bare Metal OS

This method is the classic Workstation and PC blade remoting architecture (see figure 3). The client OS is installed directly on the blade hardware and no virtualization is used. End users connect to the workstation via remote protocols such as HP RGS, Microsoft RDP, and Citrix HDX 3D via client hardware. This method is still used today by users that demand power and the performance of dedicated hardware.

Figure 3. Bare Metal GPU model (example)

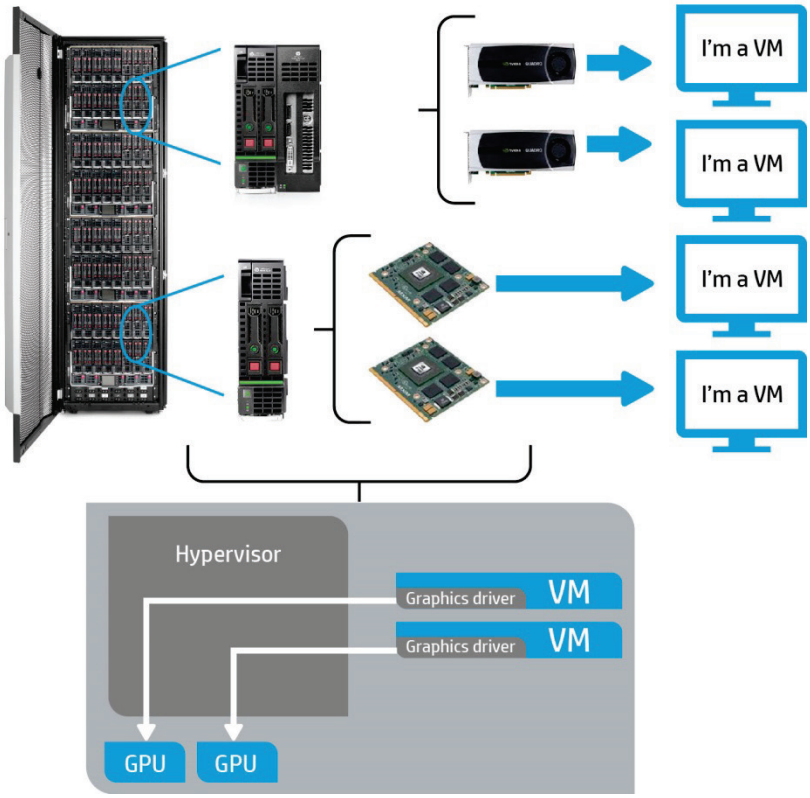

NVIDIA Quadro Mezzanine Graphics
Capture, compress, and encrypt high-end professional 3D graphics
High-end 3D professional graphics experience
Decompress and display
Unlimited distance LAN or WAN
Remote protocol
Client
Apply keyboard and mouse events
Capture keyboard and mouse events

[1] Image shown to illustrate concept and may vary with actual product and supported graphics cards.

Pass-through GPU

Also referred to generically as “Direct Attached GPU”, or vendor specific “vDGA” (VMware) and “GPU pass-through” (Citrix). This method allows discrete PCI GPU devices to be directly mapped to a virtual machine for dedicated 1:1 use by the VM (see figure 4). The virtual machine has full and direct access to GPU, including the native graphics driver, allowing for full workstation class graphics and GPU compute performance in a virtual machine. Typically intended for high end 3D and GPU Compute users, the GPU device is directly owned and managed by the VM operating systems just as in a desktop workstation. The GPU driver is loaded within the virtual machine.

Figure 4. Pass-through GPU model



Enterprise hypervisors using this technology include:

- Citrix XenServer 6.0 and newer
- VMware vSphere 5.1 and newer

Advantages:

- Up to eight workstation class VM’s per host using HP WS460c Gen8 Graphics Server Blade with HP Multi-GPU Carrier Card.
- Support for all 3D technologies including DirectX 9/10/11, OpenGL 4.3, and NVIDIA CUDA via the native NVIDIA driver in the VM.
- Graphics driver resides in the VM, enabling the virtual machine to have full and direct access to the GPU, including the native graphics driver, for full workstation performance.
- Can mix accelerated VMs and non-accelerated on same host to maximize resource utilization.

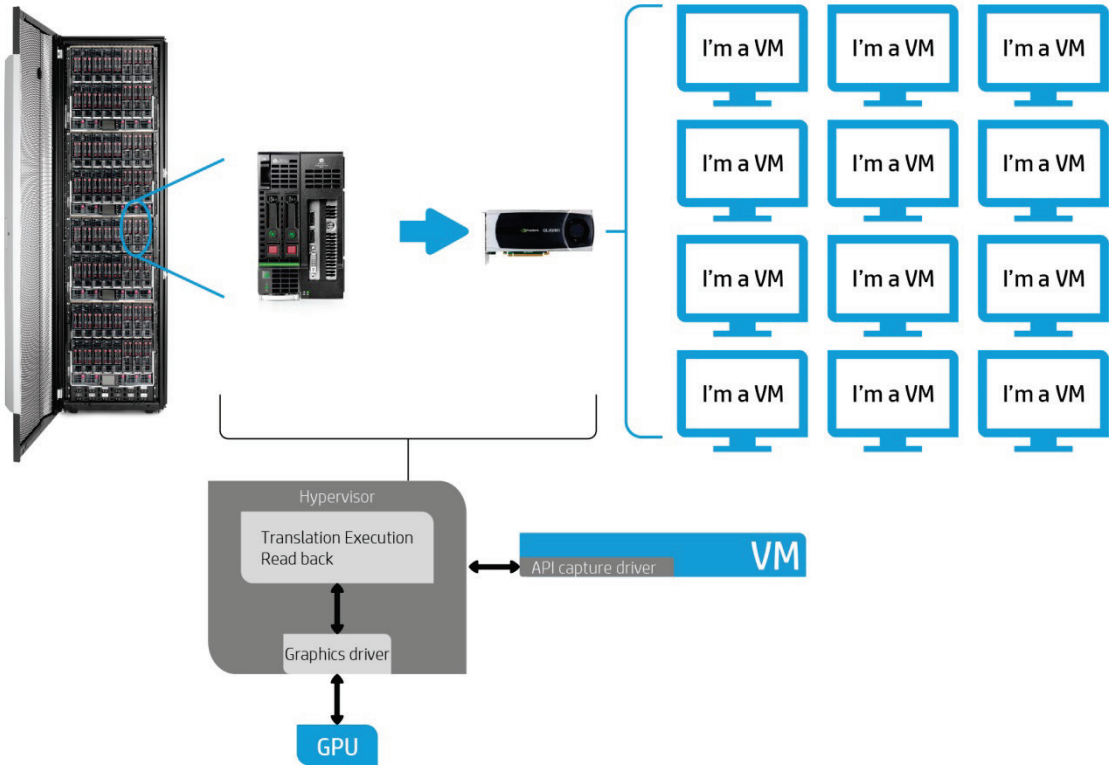
Disadvantages:

- Higher cost of ownership per connection as it has a dedicated GPU per virtual machine.
- Lower VM density per host when compared to a software virtualized GPU or non 3D desktop virtualization environment.

Software Virtualized GPU

Also referred to generically as "Shared GPU", "API intercept model", or vendor specific of "vSGA" (VMware), and "vGPU" (Microsoft RemoteFX). This method uses an API intercept model where the GPU is owned and managed by the hypervisor and all incoming graphics API requests from the VM's are intercepted via the API capture driver in the VM and redirected to and executed by the hypervisor and then sent back to VM (see figure 5). The VM does not have direct access to the GPU, and the GPU driver is loaded within the hypervisor.

Figure 5. Software Virtualized GPU model



Enterprise Hypervisors and Servers using this technology include:

- Microsoft RemoteFX vGPU
- VMware vSGA

Advantages:

- Scalability of 50+ users per GPU depending on work load.
- Load balance between multiple installed cards.
- Low cost of ownership when compared to all the accelerated graphics for desktop virtualization technologies, with highest density of VM's per GPU.
- Allows each user to have power user performance with enhanced support for DirectX 3D and Windows® Aero.
- Allows users to have "Just like desktop PC" feel and functionality (Windows Aero).

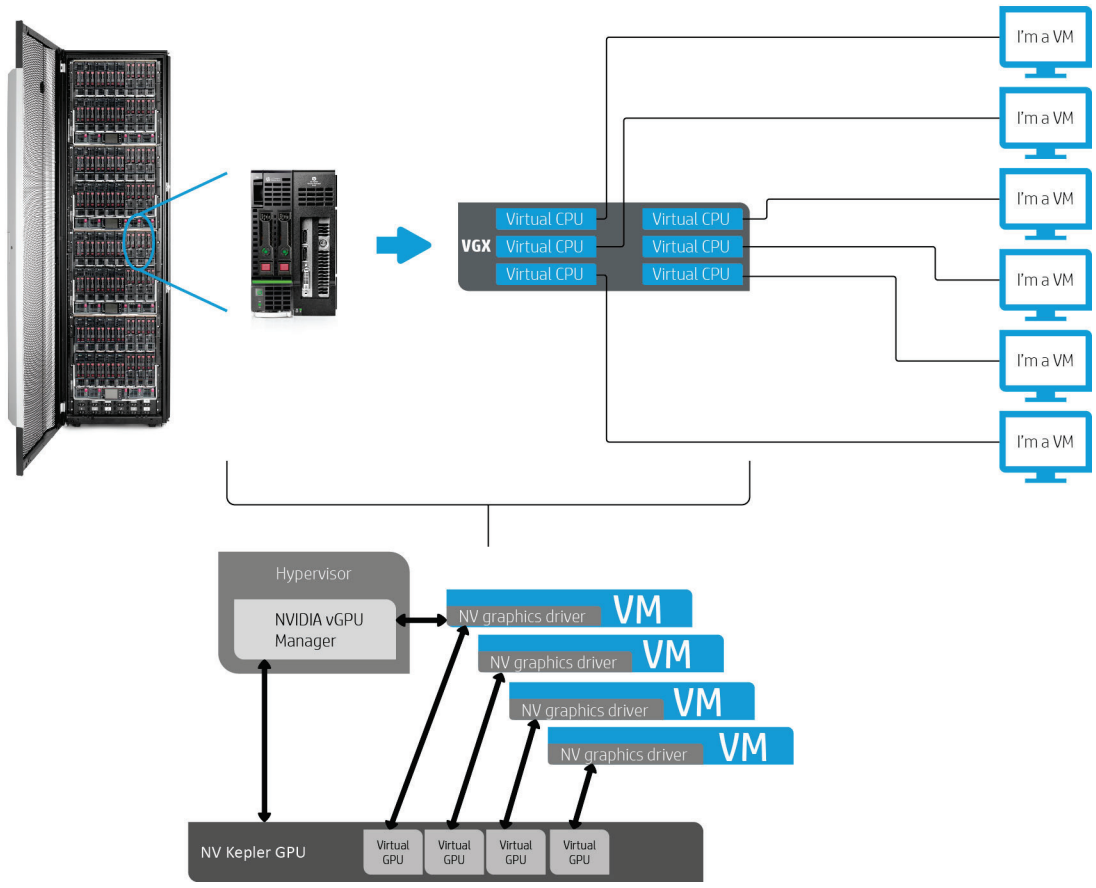
Disadvantages:

- Application compatibility due to limitation of 3D API's supported.
  - OpenGL supported versions limited.
  - DirectX supported versions may be limited to DirectX 9 in some cases.
    - GPU can become a performance bottleneck as many users draw on resource of one card and its possible that one VM can consume the resource of the GPU, for example, if it is running continuous DirectX intensive program.
    - May be unacceptable performance for high end 3D knowledge or workstation user.

Hardware Virtualized GPU—True virtual GPU

Also known as "NVIDIA GRID vGPU" is the NVIDIA/Citrix implementation of the technology, True Virtual GPU offers the benefit of GPU scaling like the software virtualized GPU (API intercept) and gives the performance of a native NVIDIA graphics driver like the pass-through models (see figure 6). This technology is currently implemented by the NVIDIA GRID K1 and K2 products. The GRID GPU is shared between multiple VM's similar to API intercept, however in this model each VM has direct access to the GPU via dedicated channels managed by the NVIDIA GRID vGPU Manager. Unlike the software virtualized GPU (API intercept) model, the NVIDIA GRID vGPU Manager within the host hypervisor manages the VM to GPU channels, guaranteeing each VM has a dedicated amount of vRAM per user and direct access to the GPU. Administrators will have ability to assign 1 to 8 users per physical GPU depending on their workload needs.

Figure 6. Hardware Virtualized GPU model



Enterprise Hypervisors using this technology include:

- Citrix XenServer with NVIDIA GRID cards and GRID vGPU technology (Not released at time of this writing)

Advantages:

- Benefit from best of both technologies of software virtualized and pass-through GPU technologies in that VM shares the resource of a GPU, but has direct access via a native graphics driver and dedicated resources of the GPU.
- Support for all 3D technologies including DirectX 9/10/11, OpenGL 4.3, and NVIDIA CUDA via the native NVIDIA driver in the VM.
- Graphics driver resides in the VM, enabling the virtual machine to have full and direct access to the GPU, including the native graphics driver, for full workstation performance.
- Lower cost of ownership than GPU pass-through in that multiple virtual machines benefit from the resource of a single GPU. Configurable GPU resource level per VM.

Disadvantages:

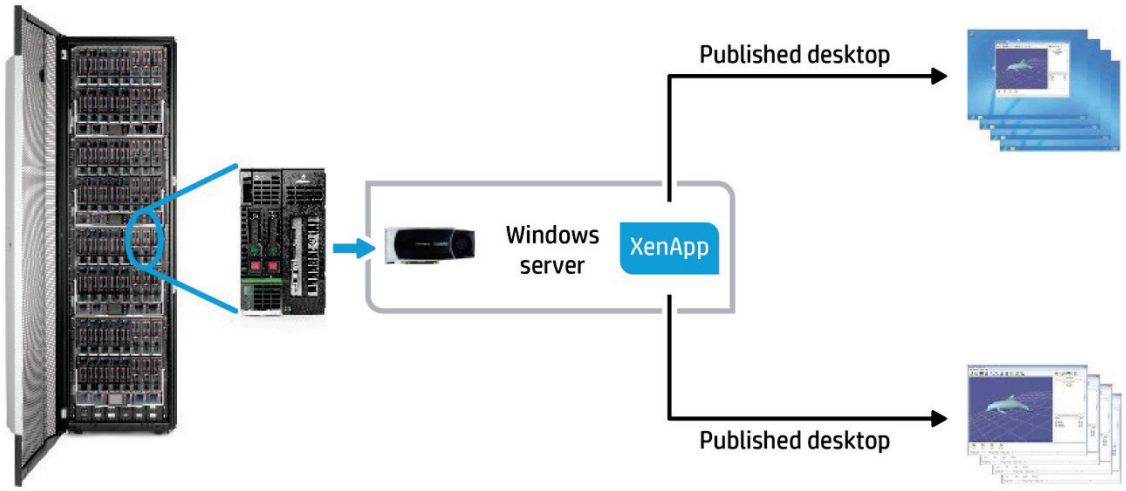
- Potential lower overall VM density per GPU as compared to software virtualized GPU model (depends on applications and workload).

Graphics accelerated desktop sessions and application virtualization

Hosted shared desktops provide a locked down, streamlined and standardized environment with a core set of applications ideally suited for users where personalization is not needed. Supporting up to 300+ users on a single server, this model offers a significant cost savings over any other virtual desktop technology. Application virtualization allows any Windows application to be centralized and managed in the data center, hosted either on multi-user terminal servers or virtual machines, and instantly published as a service to physical and virtual desktops.

In this model, 3D applications are installed on the host Windows server and published via hosted shared desktops or as a hosted published application. Since the application is running on a server equipped with a supported 3D graphics card, the application utilizes the graphics card for 3D rendering. Until now, the limiting factor has been the ability of the remoting protocol to be able to handle DirectX and OpenGL applications and publish to end user. This is changing as vendors add this support to their solution. Figure 7 shows the conceptual structure of accelerated desktop and application publishing using Citrix XenApp.

Figure 7. Graphics accelerated session and application virtualization



Enterprise Solutions supported 3D Hosted and Desktops:

- Citrix XenApp w/DirectX Hosted Apps and Desktops
- Citrix XenApp w/OpenGL Hosted Apps and Desktops (Available as of XenDesktop 5.6 FP1 and 7)

Advantages:

- Ideally suited for task workers where personalization is not needed.
- Ability to publish hosted application with 3D OpenGL and DirectX support.
- Lowest cost of ownership when compared to other accelerated graphics for VDI technologies.

Disadvantages:

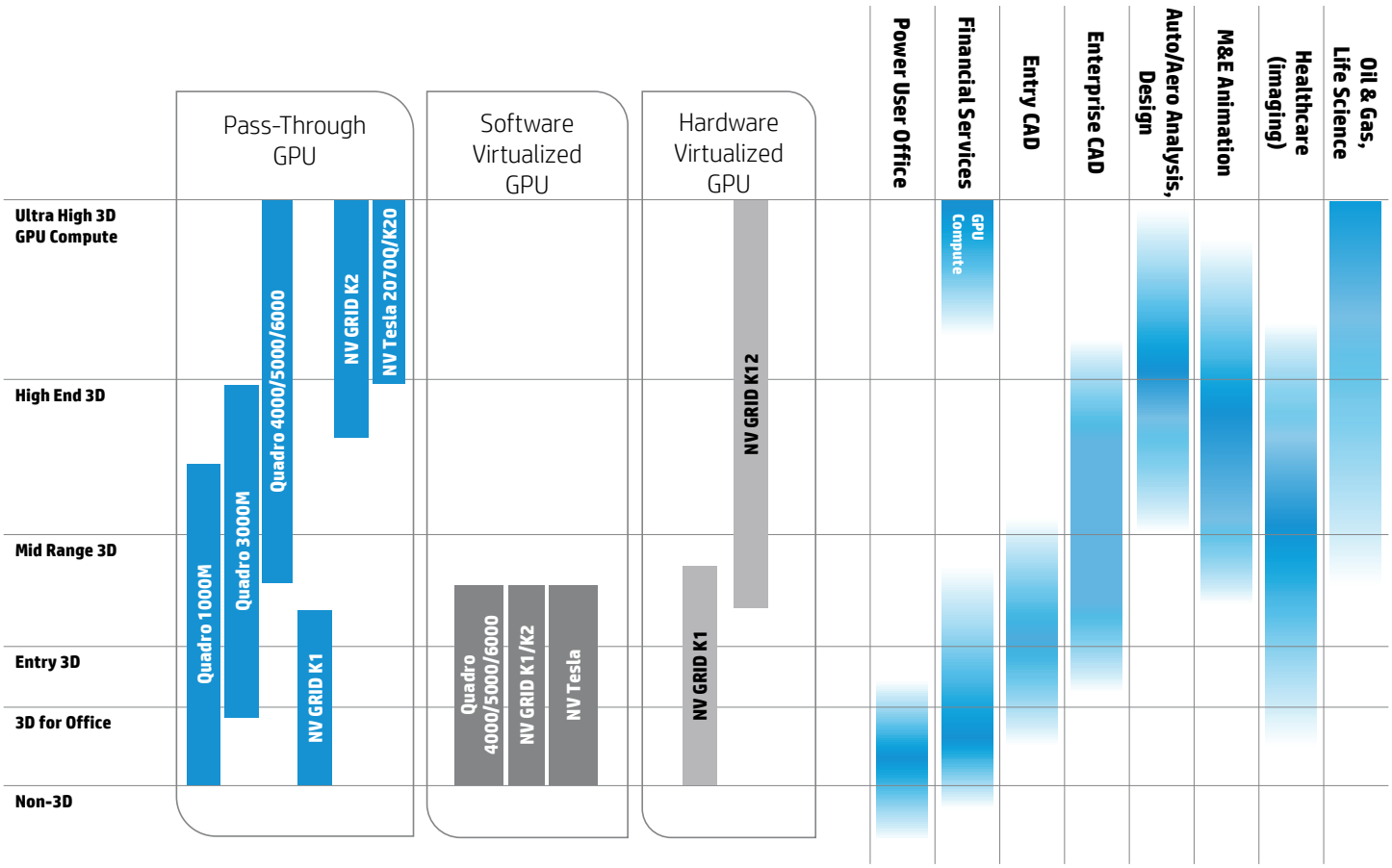
- Desktop personalization is not typically available on published desktop.
- Some 3D application may not work or be certified as a published application or multi users published desktop.
- Citrix products are the only platform with this support at the time of this writing.

Planning consideration for implementing hardware accelerated desktop virtualization technologies

Determining the right GPU and platform for your use case

The following chart (figure 8) shows comparison of GPU acceleration on desktop virtualization technologies, GPU types, and what industry segment and use case they are best fitted to.

Figure 8. Platform GPU, platform, and segment comparison



Graphics accelerated desktop virtualization platform feature comparison

Table 2a. Platform GPU feature comparison

	Bare Metal w/RGS	Citrix Multi-GPU pass-through w/HDX 3D Pro	VMware vDGA w/PCoIP	VMware vSGA w/PCoIP	Microsoft API Intercept w/RemoteFX	True Virtual GPU
Shared GPU between multiple VM's	N/A	N/A	No	Yes	Yes	Yes
GPU pass-through technology	N/A	Yes	Yes	No	No	Yes
Dual Monitor	Yes	Yes	Yes	Yes	Yes	Yes
Quad Monitor	Yes	Yes**	No	No	Yes	Yes
8 Way Monitor	No	No	No	No	Yes*	No
DirectX 9	Yes	Yes	Yes	Yes	Yes	Yes
DirectX 10	Yes	Yes	Yes	No	Yes*	Yes
DirectX 11	Yes	Yes	Yes	No	Yes*	Yes
OpenGL 2.x	Yes	Yes	Yes	Yes	Yes	Yes
OpenGL 3.x	Yes	Yes	Yes	No	No	Yes
OpenGL 4.x	Yes	Yes	Yes	No	No	Yes

* Requires Windows Server 2012 w/RemoteFX and Win8 or Win7 with RDP 8 updates.
** Requires XenDesktop 7.

Table 2b. Client OS Support Matrix for WS460c Graphics Accelerated Desktop Virtualization

	XP 32	XP 64	Vista 32	Vista 64	Win7 32	Win7 64	Win8 32	Win8 64	Server 2008 R2	Server 2012	RHEL 5	RHEL 6
WS460c G6 Bare Metal	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes**	Yes**	Yes	Yes
WS460c Gen8 Bare Metal	No	No	No	No	No	Yes	No	No	Yes**	Yes**	Yes	Yes
VMware vSGA VM	No	No	No	No	Yes	Yes	Yes	Yes	No	No	No	No
VMware vDGA VM	No	No	No	No	Yes	Yes	Yes	Yes	No	No	Yes***	Yes***
Citrix GPU pass-through VM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes*	Yes*	No***	No***
True virtual GPU VM	No	No	No	No	Yes	Yes	Yes	Yes	TBD	TBD	TBD	TBD

* GPU supported referenced in this table refers to Microsoft Hyper-V and Citrix XenApp.
** Server 2012 only supported on WS460c Gen8 when using NVIDIA GRID cards. Windows 2008/2012 Server OS only supported on WS460c as Hyper-V server.
*** Concerning Linux pass-through GPU support
- VMware only supports vDGA on Linux for DirectPath I/O (GPU compute)
- Citrix does not formally support GPU pass-through on Linux
- Neither VMware View nor Citrix XenDesktop support brokering Linux desktop at this time

Calculating VM/user density

In order to calculate how many virtual machines per GPU, you need to take into consideration the following:

- Graphics performance needed for users
- Number of monitors needed per user
- GPU used

XenServer Multi-GPU pass-through considerations

- GPU pass-through technology allows for maximum graphical performance (Workstation grade) as each VM has a dedicated GPU attached.
- XenServer only supports passing through only 1 GPU per VM.
- When you direct map graphics to a virtual machine, giving it a workstation class graphics performance, you will also need to give it workstation class resources.
  - Minimum Recommendation for best performance
    - Memory—2 GB or more (Static Assignment)
    - vCPU—Two virtual cores or more
- Once you have defined the resource needs for your "virtual workstation", you can define the remaining resources available for virtual machines running standard desktop virtualization session to maximize the resource utilization of each host server.
- GPU to VM density determined by number of GPU's, this model uses one GPU per VM, no sharing.

Hyper-V RemoteFX Considerations

- The GPU has a dedicated amount of video RAM, for example, an NVIDIA Quadro 6000 has 6 GB of video RAM. Microsoft RemoteFX (VM with vGPU connected) virtual machines consume a specific amount of video RAM based on the max number of monitors and resolution set for each virtual machine. This will dictate the maximum number of virtual machine per physical GPU. Tables 3 and 4 give you the memory allocation based on monitor and resolution configuration.
- RemoteFX uses software virtualized GPU API intercept technology, allowing multiple virtual machines to use the recourse of that GPU. The more virtual machines you configure with RFX per host, the lower the potential performance will be. For example, if 10 VM's are configured with vGPU, one of those VM's can potentially consume the GPU resource if allowed to run a high intensity 3D application, like a 3D DirectX game.
- Adding multiple physical graphics cards enhances performance and scalability as Hyper-V will load balance between cards as virtual machines start up.

Table 3. 2012 Hyper-V vRAM usage

Resolution	1 monitor	2 monitors	3 monitors	4 monitors
1024 x 768	48 MB	52 MB	58 MB	70 MB
1280 x 1024	80 MB	85 MB	95 MB	115 MB
1600 x 1200	120 MB	126 MB	142 MB	Not supported
1920 x 1200	142 MB	150 MB	168 MB	Not supported
2560 x 1600	252 MB	268 MB	Not supported	Not supported

Table 4. 2008 Hyper-V vRAM usage

Resolution	1 monitor	2 monitors	4 monitors	8 monitors
1024 x 768	75 MB	105 MB	135 MB	165 MB
1280 x 1024	125 MB	175 MB	225 MB	275 MB
1600 x 1200	184 MB	257 MB	330 MB	Not supported
1920 x 1200	220 MB	308 MB	Not supported	Not supported
2560 x 1600	Not supported	Not supported	Not supported	Not supported

VMware vDGA density considerations

- VMware vDGA pass-through GPU technology allows for maximum graphical performance (Workstation grade) as each VM has a dedicated GPU attached.
- When you direct map graphics to a virtual machine, giving it a workstation class graphics performance, you will also need to give it workstation class resources.
  - Minimum Recommendation for best performance
    - Memory—2 GB or more (Static Reserved)
    - CPU—2 virtual cores per virtual machine or more
- Once you have defined the resource needs for your "virtual workstation", you can define the remaining resources available for non-graphics accelerated desktop virtualization session to maximize the resource utilization of each host server.
- If more than one GPU is installed in a single host, any GPU's not configured for vDGA will be used for vSGA is configured.
- GPU to VM density determined by number of GPU's, this model uses one GPU per VM, no sharing.

VMware vSGA VM density considerations

- When using a GPU with vSGA, for each VM, 50% of the vRAM allocated comes from Host RAM and 50% comes from GPU.
  - How many VMs can run on host GPU/s depends on policy set for that pool.
    - If Hardware
      - If GPU resources are available it will go to next available GPU resource.
      - If no GPU resource is available it will fail to power on.
    - If Software
      - All Software
    - If Automatic
      - If GPU resources are available it will go to next available GPU resource.
      - If no GPU resource is available it will go to software.
- Adding multiple physical graphics cards enhances performance and scalability as VMware will load balance between cards as virtual machines start up.

HP WS460c Gen8 Graphics Server Blade configuration and support guidelines

The following section covers only configuration requirements specific to the WS460c and HP BladeSystem infrastructure, it does not cover the full hypervisor configuration. Refer to vender for documentation on operating system and hypervisor setup and configuration not detailed in this document.

HP WS460c Gen8 Graphics Server Blade documentation

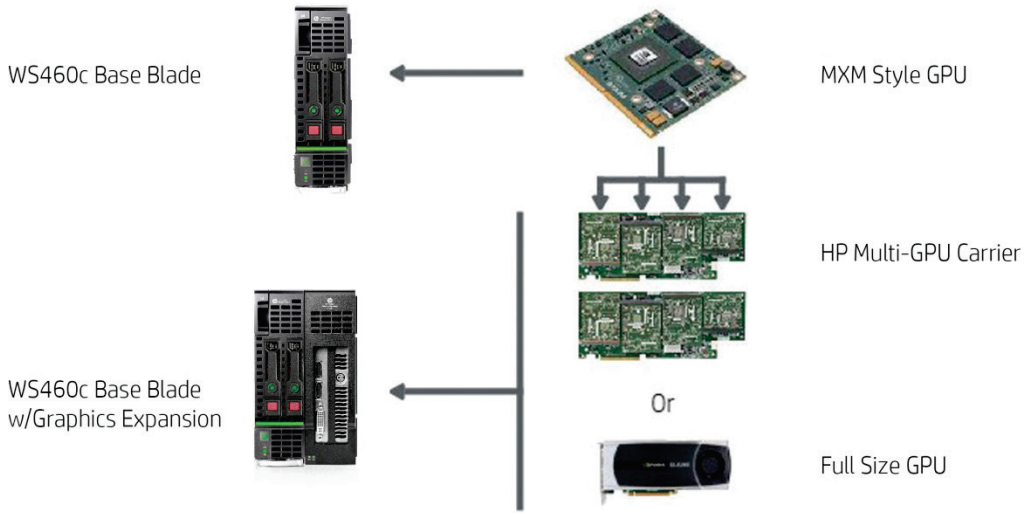
Instruction on setup and operation of the WS460c Gen8 Graphics Server Blade can be found at:
http://h20000.www2.hp.com/bizsupport/TechSupport/DocumentIndex.jsp?lang=en&cc=us&contentType=SupportManual&prodTypeId=3709945&prodSeriesId=5249678&docIndexId=64255&printver=true

Instruction on setup and operation of the WS460c G6 Graphics Server Blade can be found at:
http://h20000.www2.hp.com/bizsupport/TechSupport/DocumentIndex.jsp?lang=en&cc=us&contentType=SupportManual&prodTypeId=3709945&prodSeriesId=4012659

Understanding WS460c Gen8 Graphics Server Blade Core configurations and options

There are two ProLiant C Class generations of the WS460c, Generation 6 and 8 (G6, Gen8). There are two options for the base blade configuration (see figure 9), the single wide base blade or double wide blade with graphics expansion. The base blade supports up to two MXM style graphics cards installed on the blade mezzanine slots, while the expansion blade allows for full size high end graphics cards to be installed. Newly introduced, the WS460c Gen8 HP Multi-GPU Carrier Card allows for up to 8 GPU's (MXM Style) to be installed in the blade, creating the highest GPU density on a blade in the industry.

Figure 9. WS460c Gen8 configuration options



WS460c Graphics Server Blade OS, hardware, and firmware guidelines and support

The following information complements the Server Blade documentation with configuration requirements specific to the WS460c G6 being used for hardware accelerated graphics for desktop virtualization.

Table 5. HP WS460c G6/Gen8 infrastructure firmware recommended minimum revisions

Components	Version
HP Onboard Administrator	3.71
HP Virtual Connect	4.01
HP WS460c G6 System ROM	12/2/11
HP WS460c Gen8 System ROM	5/1/13
HP Integrated Lights-Out 2 (G6)	2.05
HP Integrated Lights-Out 4 (Gen8)	1.2

Table 6a. HP WS460c Gen8 enclosure interconnect support matrix

Interconnect	Bare Metal Client OS	Server or Hypervisor
HP 6125G Ethernet Blade Switch	Yes	Yes
HP 6125G & G/XG Ethernet Blade Switch	Yes	Yes
HP 6120G & G/XG Blade Switch	Yes	Yes
HP GbE2c Layer 2/3 Ethernet Blade Switch	Yes	Yes
Cisco Catalyst Blade Switch 3020	Yes	Yes
Cisco Catalyst Blade Switch 3120	Yes	Yes
Cisco Fabric Extender for HP BladeSystem	Yes	Yes
HP 1/10Gb Virtual Connect Ethernet Module	Yes	Yes
HP Virtual Connect Flex-10/10D Ethernet Module	Yes	Yes
HP Virtual Connect Flex-10 10Gb Ethernet Module	Yes	Yes
HP Virtual Connect FlexFabric 10Gb/24-Port	Yes*	Yes
HP Virtual Connect 8Gb 24-Port Fibre Channel	No	Yes
HP Virtual Connect 8Gb 20-Port Fibre Channel	No	Yes
Cisco MDS 8Gb Fabric Switch	No	Yes
Brocade 8Gb SAN Switch	No	Yes
HP InfiniBand for BladeSystem	No	Yes
HP 10GbE Pass-Through Module	Yes	Yes
HP 1GbE Pass-Through Module	Yes	Yes

* Supports FlexFabric as an interconnect with 530FLB with Flex-10 functionality. No support for 554M/FLB on Bare Metal.

Table 6b. HP WS460c Gen8 BLOM, NIC, and HBA support matrix

	Bare Metal OS	Server or Hypervisor
HP Flex-10 10Gb 2-port 530M Adapter	No	Yes*
HP Flex-10 10Gb 2-port 552M Adapter	No	Yes*
HP FlexFabric 10Gb 2-port 554M Adapter	No	Yes*
HP Flex-10 10Gb 2-port 530FLB Adapter	No	Yes
HP Flex-10 10Gb 2-port 530FLB Adapter	Yes	Yes
HP FlexFabric 10Gb 2-port 554FLB Adapter	No	Yes
HP Ethernet 10Gb 2-port 560FLB Adapter	No	Yes
HP LPe1205A 8Gb Fibre Channel Host Bus Adapter	No	Yes*
HP QMH2572 8Gb Fibre Channel Host Bus Adapter	No	Yes*

* Installing an additional mezzanine adaptor in WS460c will limit the number of graphics cards able to be installed.

Table 7. WS460c Supported Bare Metal Operating System (Installed directly on blade, no virtualization)

Hardware Configuration	Bare Metal XP 32	Bare Metal XP 64	Bare Metal Vista 32	Bare Metal Vista 64	Bare Metal Win7 32	Bare Metal Win7 64	VMware vSphere	Citrix XenServer	Microsoft 2008 R2 for RemoteFX or XenApp	Microsoft 2012 for RemoteFX or XenApp	RHEL 5	RHEL 6
WS460c G6 with MXM GPU	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
WS460c G6 w/Graphics Expansion**	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
WS460c Gen8 MXM GPU	No	No	No	No	No	Yes	Yes	Yes	Yes	No*	Yes	Yes
WS460c Gen8 w/Graphics Expansion**	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes*	Yes	Yes
WS460c Gen8 w/Multi-GPU Carrier	No	No	No	No	No	No	Yes	Yes	No	No	No	No

* Only supported NVIDIA cards for RemoteFX on Server 2012 are NVIDIA K1 and K2 full size cards.
** HP WS460c w/Graphics Expansion allows for all full size card and HP Multi-GPU carrier card to be installed.

Table 8. WS460c Gen8 maximum supported graphics cards per platform/configuration

Hardware Configuration	Bare Metal	Microsoft 2008 R2 for RemoteFX and XenApp	Microsoft 2012 for RemoteFX and XenApp	VMware vDGA	VMware vSGA	Citrix GPU Pass-Through	True Virtual GPU***
NV Quadro K4000****	Max 1*	Max 1*	Max 1*	Max 1*	No	Max 1*	No
NV Quadro K5000	Max 1*	Max 1*	Max 1*	Max 1*	No	Max 1*	No
NV Quadro K6000****	Max 1*	Max 1*	Max 1*	Max 1*	No	Max 1*	No
NV Quadro K20****	No	No	No	Max 1*	No	No	No
NV Quadro 5000	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	No
NV Quadro 6000	Max 1*	Max 1*	Max 1	Max 1*	Max 1*	Max 1*	No
NV Quadro 3000M	Max 1	Max 1	No	Max 6**	No	Max 6**	No
NV Quadro 1000M	Max 2	No	No	Max 8**	No	Max 8**	No
NV Quadro 500M	Max 2	No	No	No	No	No	No
NV GRID K1****	No	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	Planned
NV GRID K2****	No	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	Planned
NV Quadro Tesla M2070Q	No	No	No	Max 1*	Max 1*	No	No
NV Quadro 3800	No	No	No	No	No	No	No
NV Quadro 4800	No	No	No	No	No	No	No
NV Quadro 5800	No	No	No	No	No	No	No
NV Quadro 880M	No	No	No	No	No	Max 1****	No
NV Quadro 2800M	No	No	No	Max 1****	No	Max 1****	No

* Requires Graphics Expansion Blade.
** Requires HP Multi-GPU Carrier Card.
*** "True Virtual GPU" is a reference to a future technology and not the name of a product.
**** These cards are in plan and not released at the time of this writing.

Table 9. WS460c G6 maximum supported graphics cards per platform/configuration

Hardware Configuration	Bare Metal	Microsoft 2008 R2 for RemoteFX or XenApp	Microsoft 2012 for RemoteFX or XenApp	VMware vDGA	VMware vSGA	Citrix GPU Pass-Through	True Virtual GPU***
NV Quadro 5000	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	No
NV Quadro 6000	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	No
NV Quadro 3000M	No	No	No	No	No	No	No
NV Quadro 1000M	No	No	No	No	No	No	No
NV Quadro 500M	No	No	No	No	No	No	No
NV Quadro Tesla M2070Q	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	Max 1*	No
NV Quadro 3800	Max 1	No	No	No	No	No	No
NV Quadro 4800	Max 1	No	No	No	No	No	No
NV Quadro 5800	Max 1	No	No	No	No	No	No
NV Quadro 880M	Max 2	Max 2	Max 2	No	No	No	No
NV Quadro 2800M	Max 1	Max 1	Max 2	No	No	No	No

* Requires Graphics Expansion Blade.
** Requires HP Multi-GPU Carrier Card.
*** "True Virtual GPU" is a reference to a future technology and not the name of a product.
**** There cards are announced but not released at the time of this writing.

For latest listing of supported cards for each server OS and Hypervisor platform, please visit the following sites:

Citrix—hcl.xensource.com/

Microsoft—windowsservercatalog.com

VMware—vmware.com/resources/compatibility

Required WS460c BIOS configurations

Configuring Server Blade video mode

The WS460c G6/Gen8 Graphics Server Blade has four distinct graphics modes that may be available. Depending on what OS is installed, different modes are used. In a nutshell, the Graphics Server Blade has an embedded graphics card as well a high end graphics card, and in some configuration, it is necessary to disable one or the other card. Table 10 shows the supported modes for each environment and the following sections describe each.

Table 10. Proper use of video mode for each operating system

Hardware Configuration	Bare Metal	Microsoft Server for RemoteFX and XenApp	Citrix XenServer	VMware vSphere
OS setup and configuration	Setup mode	Setup mode	Setup mode	Setup mode
Running in production	User mode	User mode	Setup mode	Setup mode

User mode

- This mode is used when it is necessary to disable the embedded graphics card because the installed operating does not support using two different graphics architectures.
- In this mode the add-in card is enabled and the embedded is disabled. The OS only sees the add-in card.
- This is the primary production mode when the following operating systems are installed:
  - Bare Metal Installations of Windows or Linux
  - Microsoft RemoteFX environments
  - Microsoft/XenApp environments
- In this mode the iLO and front SUV console is available in POST but when control is passed to the OS, it becomes inaccessible because it uses the embedded video card to generate its video. In this mode the server console will show a message indicating it is in User mode.
- In this mode, once the operating system is booted, the system can be accessed via remote protocol only.

Setup mode

- This mode is used in the following configurations:
  - Used in full production mode when the installed operating does support using two different graphics architectures at the same time. Currently the operating systems that use this mode for production include the VMware and XenServer Hypervisors.
  - Used during system install and configuration when the installed operating does not support using two different graphics architectures at the same time. For example, this mode is used during install, configuration and driver install for Windows 7 and Server, as well as Linux (Bare Metal). The mode is switch to “User” before going into production.
- This mode has both video cards enabled but the add-in card is secondary and the embedded is primary.
- This is the primary production mode when the following operating systems are install:
  - Citrix XenServer
  - VMware vSphere
- In this mode the iLO console is accessible at all times.

Admin mode

- This mode is used for troubleshooting only and is not used in production
- This mode disables the add-in card and only the ATI is active

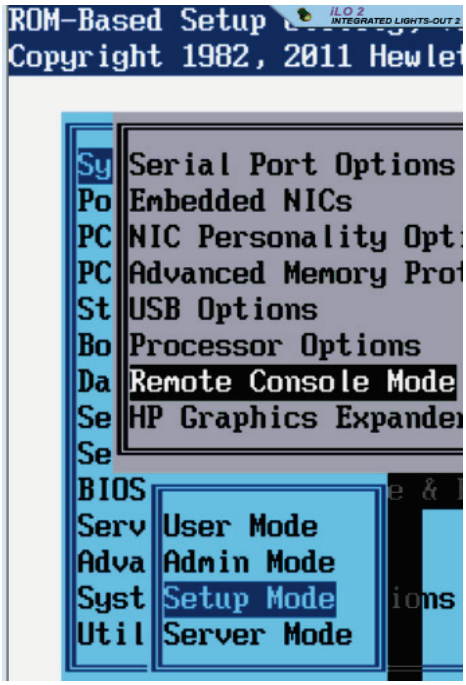
Server mode (WS460c G6 only)

- This mode is deprecated and rarely used for troubleshooting only and is not used/supported in production
- This mode operates in the same way as Admin mode

Procedure to set Remote Console:

1. Using either iLO remote console or the Local I/O Connector, view the boot console using iLO 2, which provides direct control of the Graphics Server Blade.
2. When prompted during boot, press the F9 key. The ROM-based Setup Utility appears.
3. Select System Options > Remote Console Mode. The current Remote Console Mode appears. (figure 10)
4. To change the Remote Console Mode, press Enter. The Remote Console Mode menu appears. Use the Up and Down arrow keys to select the desired mode. When done, press Enter and then perform the steps indicated to exit the ROM-based Setup Utility.
5. The Graphics Server Blade will reboot, and then the Remote Console Mode will be in effect.

Figure 10. Remote Console Mode



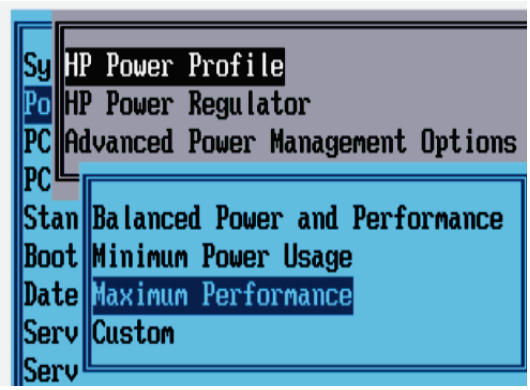
Setting the Graphics Server Blade to Static High Performance

When used with hypervisors, the Graphics Server Blade must be in Static High Performance mode. A BIOS setting dictates the performance mode. This setting must be manually set.

Configuring the Graphics Server Blade performance mode. Procedure to set performance mode:

1. Using either iLO remote console or the Local I/O Connector, view the boot console using iLO 2, which provides direct control of the Graphics Server Blade.
2. When prompted during boot, press the F9 key. The ROM-based Setup Utility appears.
3. As shown in figure 11, select Power Management Options > HP Power Profile.
4. To change mode, press Enter. The mode menu appears. Use the Up and Down arrow keys to select the desired mode.
5. Select and enable “Maximum performance.”

Figure 11. High Performance Mode



Configuring the Graphics Server Blade PCIe mode
(Only applies to WS460c G6)

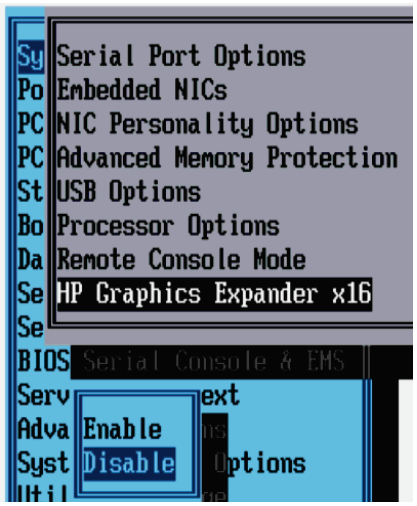
Setting the HP Graphics Expander BIOS settings (Only applies to WS460c G6). The WS460c G6 supports up to two full size graphics cards in the graphics expansion bay. The PCIe graphics expansion bay slots can be configured for one card at x16 or two cards at x8. A BIOS setting dictates if the system can see one or two cards. This setting must be manually set. Setting "HP Graphics Expander x16"

Enabled
- This mode enables one slot at x16

Disabled
- This mode enabled two slots at x8

Procedure to set PCIe mode:
1. Using either iLO remote console or the Local I/O Connector, view the boot console using iLO 2, which provides direct control of the Graphics Server Blade.
2. When prompted during boot, press the F9 key. The ROM-based Setup Utility appears.
3. As shown in figure 12, select System Options > HP Graphics Expander x16. The current Expander mode appears.
4. To change mode, press Enter. The mode menu appears. Use the Up and Down arrow keys to select the desired mod.
   A. Enable—Turns on one card at x16
   B. Disable—Turns on two cards at x8
5. When done, press Enter, and then perform the steps indicated to exit the ROM-based Setup Utility.
6. The Graphics Server Blade performs a reboot, and then the Remote Console Mode appears.

Figure 12. PCIe Mode

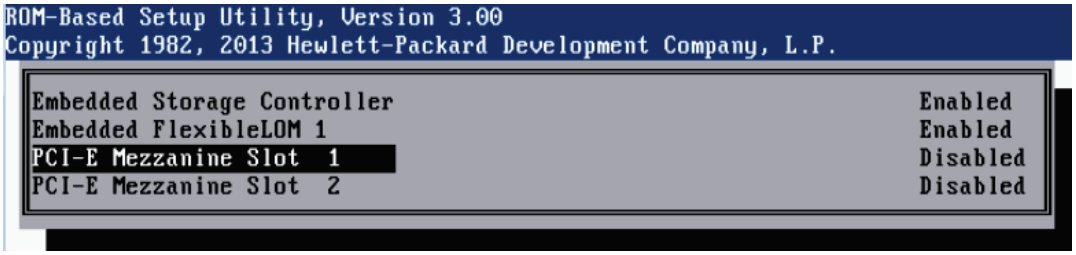


Turn off PCIe Gen3 support
(Only Needed on WS460c Gen8 with NVIDIA K4000/K5000/K6000)

Although this is a default setting in WS460c Gen8 Graphics Server system BIOS as of 5/1/13 release, if this BIOS setting is enabled when using the NVIDIA K4000/5000/6000, the system will NMI on start up. To prevent this, disable Gen3 support at following location.

Procedure to set PCIe mode:
1. Access BIOS and go to Power mgmt. options >> Adv. Power mgmt. options >> PCIe Gen 3 Control
2. Set both PCIe slots to disable

Figure 13. PCIe Gen3 support


ROM-Based Setup Utility, Version 3.00
Copyright 1982, 2013 Hewlett-Packard Development Company, L.P.
Embedded Storage Controller Enabled
Embedded FlexibleLOM 1 Enabled
PCI-E Mezzanine Slot 1 Disabled
PCI-E Mezzanine Slot 2 Disabled

Platform specific version and configuration recommendations

The following section covers configuration requirements and recommendations specific to the desktop virtualization platform installed. These recommendations are the minimum version as recommend by HP for use on HP WS460c Gen8 Graphics Server Blade. These requirements may not reflect the vendor specific minimum required version recommendations. It does not cover the full hypervisor configuration. Refer to vender for documentation on operating system and hypervisor setup and configuration not detailed in this document.

Citrix XenServer Direct Map system configuration recommendations

Table 11. Required BIOS setting for XenServer 6 on Graphics Server Blade

Setting	Value
Advanced Options/Option ROM Loading Sequence (G6 Blades Only)	Load Option Card Devices First
Power Management Options/HP Power Profile (see section: Setting the Graphics Server Blade to Static High Performance)	Maximum performance
Video Mode for setup and production use (see section: Configuring Graphics Server Blade video mode.)	Setup mode

Table 12. HP recommended minimum Citrix versions for best user performance

Components	Software description
Citrix XenServer	Citrix XenServer 6.0.2 (6.2 preferred)
Citrix XenDesktop Controller	XenDesktop 5.6 FP1 (XD 7 preferred)
Citrix Receiver	Citrix Receiver 3.1 with online plug-in 13.1.0.89
Citrix XenDesktop Virtual Desktop Agent	XenDesktop 5.6 FP1 VDA

XenServer 6 Important notes

- Use the latest available driver for virtual machine client OS from NVIDIA website.
- No driver for NVIDIA card is installed on XenServer, the driver is loaded directly on the virtual machine just like a desk side unit after the GPU is assigned to the VM.
- Because XenServer does not load drivers for the NVIDIA cards, the systems can run in "Setup" video mode to allow access through iLO remote consoles.
- XenServer only allows one GPU to be attached to any virtual machine.
- For best performance, run HDX 3D Citrix ICA and HDX 3D Pro on client devices with at least dual core 2 GHz processor. (Minimum requirement for a single monitor configuration is a single core 2 GHz processor.)
- Need to configure virtual monitors in the GPU control panel of the VM to support multiple monitors with Citrix XenDesktop HDX 3D Pro.

Configure Citrix VM and HDX 3D Pro for best performance

- 2 vCPU per virtual machine or more.
  - In order to set more than two cores per socket (allowing for 4 cores or more) on a virtual machine, you must configure the virtual machine to enable multiple cores per socket. Use the following command for XenServer console to enable up to 4 cores per socket and set the virtual machine to use 4 cores at next boot.

    ```
    xe vm-param-set uuid=<VM UUID> platform:cores-per-socket=4 VCPUs-max=8 VCPUs-at-startup=4
    ```
- 2 GB or more of virtual machine memory.
- Turn off Aero support unless needed. Aero will consume one CPU core.
- Smaller screen resolutions perform better.

VMware system configuration recommendations

Table 13. Required BIOS setting for Graphics Server Blade

Setting	Value
Advanced Options/Option ROM Loading Sequence (G6 Blades Only)	Load Option Card Devices First
Power Management Options/HP Power Profile (see last section)	Maximum performance
Video Mode for Setup and Production	Setup mode

Table 14. VMware required minimum versions for vSGA

Components	Software description
ESXi Host	5.1
VMware View Connection Server	5.2
VMware View Agent	5.2
View Client	5.3
NVIDIA ESXi OS Driver for (vSGA only)	320.00 or newer
NVIDIA Graphics Driver for VM (vDGA only)	320.00 or newer

VMware important notes

- Using VMware vSGA technology does not require any custom configuration on target VM other than minimum View agent version and standard View agent setup.
- Using VMware vDGA technology directly attaches a GPU to a VM for maximum graphics performance. For that reason, it is required that proper video driver be installed within the virtual machine.
- vSGA only has support for DirectX 9 and OpenGL 2.1.
- Some programs look for a recognized GPU and will not start or run properly on vSGA.
- High availability and vMotion
  - High availability and vMotion **is not** supported when using in vDGA mode.
  - High availability and vMotion **is** supported when using vSGA mode.

Configure VMware VM and PCoIP for best performance

- Configuring VM for vDGA for workstation class performance
  - Minimum 2 vCPU per virtual machine or more.
  - 2 GB or more of virtual machine memory.
- Turn off Aero support unless needed.
- Depending on workload, vDGA end clients may require more power (2 GHz processor or better).

Microsoft Hyper-V/RemoteFX system configuration recommendations

Table 15. Microsoft required minimum versions

Components	Software description
Microsoft RFX	Microsoft Windows Server 2008 R2 SP1 or Server 2012
Microsoft RDP	Microsoft RDP 7.1 or later version (Recommend RDP 8 or RDP 8 update for Windows 7)

Table 16. Required BIOS setting for 2008 and 2012 Server on Blade Workstation WS460c

Setting	Value
Advanced Options/Option ROM Loading Sequence (G6 Blade Only)	Load Option Card Devices First
Power Management Options/HP Power Profile (see last section)	Maximum performance
Video Mode for Server Setup	Setup mode
Video Mode for Production*	User mode

* Once in production (User mode) the server console will only be accessible via remote connection (see section: Configuring Workstation Blade video mode.)

What's new in Server 2012 for Accelerated Graphics for desktop virtualization?

- Improvements:
  - Adaptive Graphics.
  - Intelligent Transports.
  - Optimized Media Streaming.
  - Adaptive Network Auto Detect.
  - DirectX 11 Support with vGPU.
  - Single Sign-On.
  - Email and web discovery of Remote Applications and desktops.
  - Multi Touch.
  - USB Redirection.
  - Metro-style Remote Desktop.

For more information on these improvement, visit
http://blogs.technet.com/b/windowsserver/archive/2012/05/09/windows-server-2012-remote-desktop-services-rds.aspx

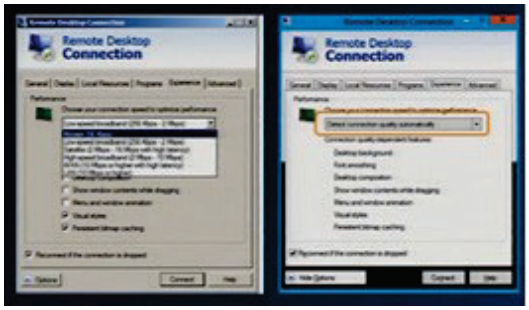
Hyper-V RemoteFX considerations

- The GPU has a dedicated amount of video RAM, for example, an NVIDIA Quadro 6000 has 6 GB of video RAM. Microsoft RemoteFX virtual machines consume a specific amount of video RAM based on the max number of monitors and resolution set for each virtual machine. This will dictate the maximum number of virtual machine per GPU. See table 3 and 4 for full details.
- RemoteFX technology shares the resources of the GPU, allowing multiple virtual machines to use the recourse of that GPU. The more virtual machines you configure with RFX per host, the lower the potential performance will be if running heaver 3D workloads.
- Adding multiple graphics cards enhances performance as Hyper-V will load balance between cards as virtual machines start up.
- Once you have defined the resource needs for your "virtual workstation", you can define the remaining resources available for virtual machines running standard VDI session to maximize the resource utilization of each host server.

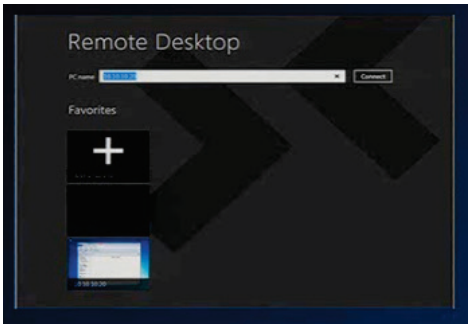
Microsoft RemoteFX important notes

- Server 2008
  - Use the latest available driver for Windows Server 2008 R2 from NVIDIA website.
    - Windows 2008 requires drivers for the NVIDIA cards installed on host, but does not support two types of video cards running at the same time. Because of this of this the following modes must be used.
      - Setup mode—for systems install, setup, and NVIDIA driver loading.
      - User mode—for production RFX mode.
        - In this mode it is not possible to reach iLO console through iLO, OA, or front I/O dongle. You must enable remote RDP console for systems management.
  - You can install more than one graphics card of same type and Windows will load balance between them at virtual machine startup.
  - GPU assignments are not dynamically managed after virtual machine startup for load balancing.
  - At this time this solution only has minimal support of OpenGL applications.
- Server 2012
  - In Server 2012, the term RemoteFX definition has been expanded to include the following features: RemoteFX for WAN, Adaptive Graphics, Media Remoting, Multi Touch, USB Redirection, Metro Style Remote Desktop App, Choice of Software or Hardware GPU, RemoteFX supported on Sessions, VM's, and Physical Machines.
  - Use the latest available driver for Windows Server 2008 R2 from NVIDIA website.
    - Windows 2012 requires drivers for the NVIDIA cards installed on host, but does not support two types of video cards running at the same time. Because of this the following modes must be used.
      - Setup mode—for systems install, setup, and NVIDIA driver loading.
      - User mode—for production RFX mode.
        - In this mode it is not possible to reach iLO console through iLO, OA, or front I/O dongle. You must enable remote RDP console for systems management.
  - GPU assignments are not dynamically managed after virtual machine startup for load balancing.

- At this time this solution only has minimal support of OpenGL applications.
- GPU is no longer required for RemoteFX as the 3D components can be rendered in software, but having a GPU will significantly improve performance and offload from CPU.
- There are now two clients that can be used in Windows 8
  - Classic RDP Client



  - New Metro RDP Client
    - Required to support
      - Touch screen
      - Integrated app publishing



Configure RemoteFX for best performance

- RemoteFX also supports two Group Policy settings that give administrators the flexibility to manually choose the best configuration for their scenario.
  - Policies are under this path: "ComputerConfiguration\AdministrativeTemplates\WindowsComponents\Remote Desktop Services\Remote Desktop Session Host\Remote Session Environment."
    - "Configure image quality for RemoteFX Adaptive Graphics."

      This policy setting specifies the visual quality for a remote session. Administrators can use this option to balance network bandwidth usage with visual quality delivered. The options are Medium (default), High, and Lossless. "Medium" quality consumes the lowest amount of bandwidth, "High" quality raises the image quality with a moderate raise in bandwidth consumption, while "Lossless" uses lossless encoding, which preserves full color integrity but requires significant increase in bandwidth. "Configure RemoteFX Adaptive Graphics."
    - Connection Speed must be set to "LAN (10 Mb/s or higher)."
    - This policy setting allows the administrator to choose the encoding configuration to be optimized for server scalability or bandwidth usage. By default RemoteFX chooses the best configuration at runtime (RDP 8 Only), and could dynamically switch between configurations based on network condition.

      For more information, see this document: http://blogs.msdn.com/b/rds/archive/2012/08/06/remotefx-adaptive-graphics-in-windows-server-2012-and-windows-8.aspx

- When connecting to RFX session from RDP 7 RDP client, ensure the following settings are configured:
  - Display Tab
    - Colors setting set to "Highest Quality (32 bit)."
  - Experience Tab
    - Connection Speed must be set to "LAN (10 Mb/s or higher)."
    - All check boxes should be checked.

Appendix A— Alternative remoting protocols and brokers

The following section covers only alternative options for remoting protocols and brokers that we have seen successfully used. The fact that these options are discussed here does not mean that they fully supported by the different platforms.

Leostream—Connection Broker
A connection broker manages the assignment and connection of end users to their virtual machines. Individual virtualization partner provide integrated connection brokers as part of their virtualization stack. In some cases, however, a third party connection broker becomes necessary to provide the flexibility and functionality required to support more complicated end-user environments.

The Leostream Connection Broker is a vendor-independent connection broker that allows you to manage a mixed environment that includes Citrix XenServer, Microsoft Hyper-V, VMware vSphere, and Graphics Server Blades in a single Administrator Web interface, and simultaneously connect users to a mixed environment from a single login screen. Leostream integrates with your existing infrastructure, including authentication servers, load balancers, and SSL VPNs, and supports all major display protocols, including Citrix HDX, Microsoft RemoteFX, and HP RGS.

Leostream Connection Broker policies allow you to control which resources the user is offered, how long they can use the resources, and what type of end-user experience they received, based on the user's identity and also on their location. The Leostream high availability features, such as backup pools, Connection Broker clustering, and database mirror-awareness, ensure that users remain productive in the event of failure as minor as a single blade to as large as an entire data center.

You may want to investigate Leostream if your environment fits into any of the following categories.

- It includes Linux remote desktops as well as Windows desktops
- You have different user groups who use RGS, HDX, RemoteFX, RDP, or other display protocols
- You have authentication servers other than Active Directory, such as eDirectory, OpenLDAP, or NIS
- You have a multi-domain environment with untrusted domains
- You have complicated use case requirements that change as the user's location changes

The Connection Broker is a virtual appliance that can be imported into a VMware, Citrix, or Microsoft virtualization layer. The Connection Broker requires virtual resources equivalent to the following hardware:

- 1500 MHz or faster Intel® Pentium® IV processor (or equivalent)
- 1 vCPU
- 2.0 GB of RAM
- 8 GB of hard drive space
- Bridged Ethernet adapter, ideally with Internet connectivity

Adding a second CPU to the virtual appliance does not improve Connection Broker performance as the appliance does not take advantage of the new CPU. To improve Connection Broker performance, build a Connection Broker cluster. A Connection Broker cluster is a group of Connection Brokers that share the same Microsoft SQL Server database. A common cluster uses three to five Connection Brokers.

For more information - leostream.com

HP Remote Graphics Software—remoting protocol
Remote workstations are breaking free of network limitations with HP Remote Graphics Software (RGS) 6.0. HP RGS is the collaboration and remote desktop solutions for serious workstation users and their most demanding applications. All applications run natively on the remote workstation and take full advantage of its graphics resources. The desktop of the remote workstation is transmitted over a standard network to a window on a local computer using advanced image compression technology specifically designed for digital imagery, text and high frame rate video applications. A local keyboard and mouse are supported as well as redirection of most USB1 devices to provide an interactive, high performance workstation experience.

Advanced Features in HP RGS 6.0:
- Advanced Video Compression for a significant reduction in network bandwidth usage
- Support for Linux environments
- Improved WAN performance with integrated HP Velocity technology

HP RGS is free on HP Personal Workstations. There are no monthly fees, the Receiver is a free download, and the Sender is free on HP Z Workstations and EliteBook Workstations. A license is required to run the HP RGS Sender on all other hardware. Only existing HP RGS customers can purchase upgrade licenses to HP RGS 6.0. For new HP RGS customers please contact your HP Account rep for more details.

Appendix B—Choosing the best graphics hardware and client virtualization solution

Multiple user per server—Up to 2 displays per user

Pass-through GPU:
Also referred to generically as "Direct Attached GPU", or vendor specific "vDGA" (VMware) and "GPU pass-through" (Citrix). This method allows discrete PCI GPU devices to be directly mapped to a virtual machine for dedicated 1:1 use by the VM

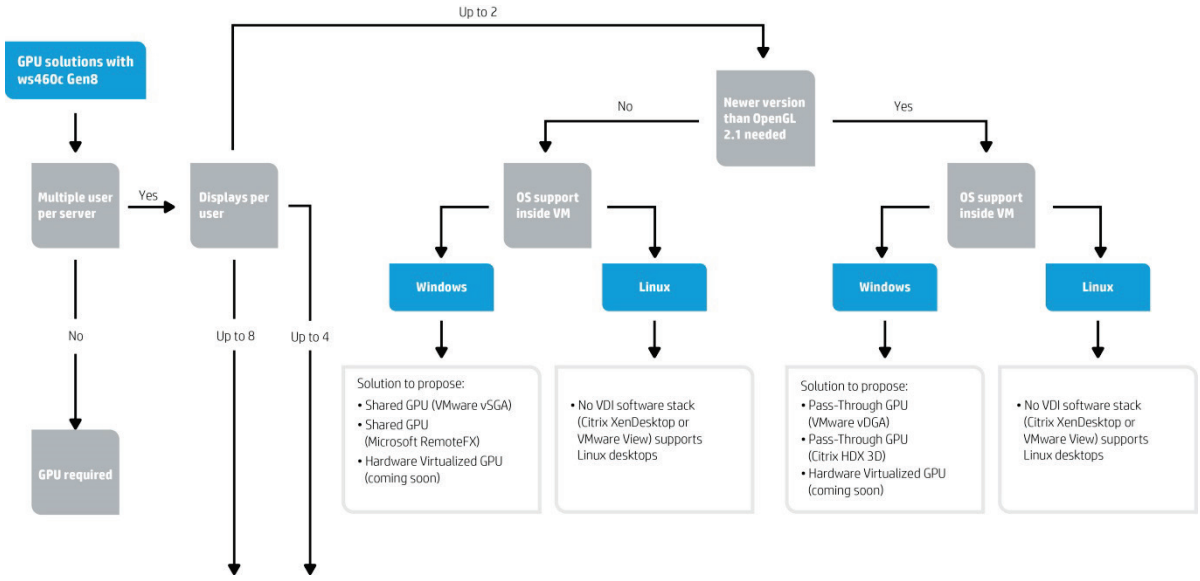
GPU: Multi GPU Carrier Board (8x Q1000m) (low to mid end user)	Density: 8 users per blade, 64 user per enclosure
GPU: Multi GPU Carrier Board (6x Q3000m) (mid to high end user)	Density: 6 users per blade, 48 user per enclosure
GPU: NVIDIA GRID K1 (low end user)	Density: 4 users per blade, 32 user per enclosure
GPU: NVIDIA GRID K2 (very high end user)	Density: 2 users per blade, 16 user per enclosure

Shared GPU:
Also referred to as "API intercept model", or vendor specific of "vSGA" (VMware), and "vGPU" Microsoft RemoteFX). This method uses an API intercept model where the GPU is owned and managed by the hypervisor and all incoming graphics API requests from the VM's are intercepted via the API capture driver in the VM and redirected to and executed by the hypervisor and then sent back to VM (see figure 5). The VM does not have direct access to the GPU, and the GPU driver is loaded within the hypervisor.

GPU: NVIDIA Quadro 5000, 6000, NVIDIA GRID K1 and K2	Density: depends on workload of each user

Hardware Virtualized GPU—True virtual GPU
Also known as "NVIDIA VGX Hypervisor" in the NVIDIA/Citrix implementation of the technology, True Virtual GPU is conceptually a hybrid of software virtualized GPU (API intercept) and pass-through models (see figure 6). This technology is currently implemented by the NVIDIA GRID K1 and K2 products. The GRID GPU is shared between multiple VM's as in API intercept, however in this model each VM has direct access to the GPU via dedicated channels managed by the NVIDIA VGX Hypervisor. Unlike the software virtualized GPU (API intercept) model, the NVIDIA hypervisor within the host hypervisor manages the VM to GPU channels, guaranteeing each VM has a dedicated amount of vRAM per user and direct access to GPU.

Figure 14. Decision tree, multiple user per server—Up to 2 displays per user



Multiple user per server—Up to 4 displays per user

Citrix GPU pass-through for 4 Displays (coming soon, no ETA date)

This method allows discrete PCI GPU devices to be directly mapped to a virtual machine for dedicated 1:1 use by the VM

GPU: NVIDIA GRID K1 (low end user) Density: Unknown at this time

GPU: NVIDIA GRID K2 (very high end user) Density: Unknown at this time

Shared GPU:

Also referred to as "API intercept model", or vendor specific of "vSGA" (VMware), and "vGPU" Microsoft RemoteFX). This method uses an API intercept model where the GPU is owned and managed by the hypervisor and all incoming graphics API requests from the VM's are intercepted via the API capture driver in the VM and redirected to and executed by the hypervisor and then sent back to VM (see figure 5). The VM does not have direct access to the GPU, and the GPU driver is loaded within the hypervisor.

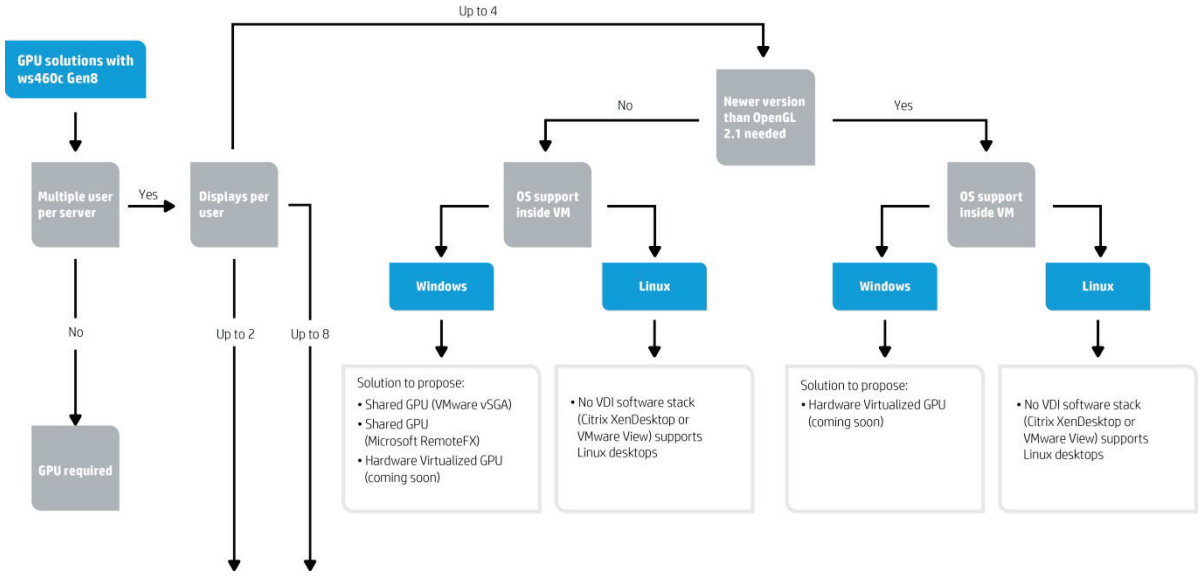
GPU: NVIDIA Quadro 5000, 6000, NVIDIA GRID K1 and K2 Density: depends on workload of each user

Hardware Virtualized GPU—True virtual GPU

Also known as "NVIDIA VGX Hypervisor" in the NVIDIA/Citrix implementation of the technology, True Virtual GPU is conceptually a hybrid of software virtualized GPU (API intercept) and pass-through models (see figure 6). This technology is currently implemented by the NVIDIA GRID K1 and K2 products. The GRID GPU is shared between multiple VM's as in API intercept, however in this model each VM has direct access to the GPU via dedicated channels managed by the NVIDIA VGX Hypervisor. Unlike the software virtualized GPU (API intercept) model, the NVIDIA hypervisor within the host hypervisor manages the VM to GPU channels, guaranteeing each VM has a dedicated amount of vRAM per user and direct access to GPU.

GPUs: NVIDIA GRID K1 and K2 Density: depends on workload of each user

Figure 15. Decision tree, multiple user per server—Up to 4 displays per user



Multiple user per server—Up to 8 displays per user

Microsoft RemoteFX

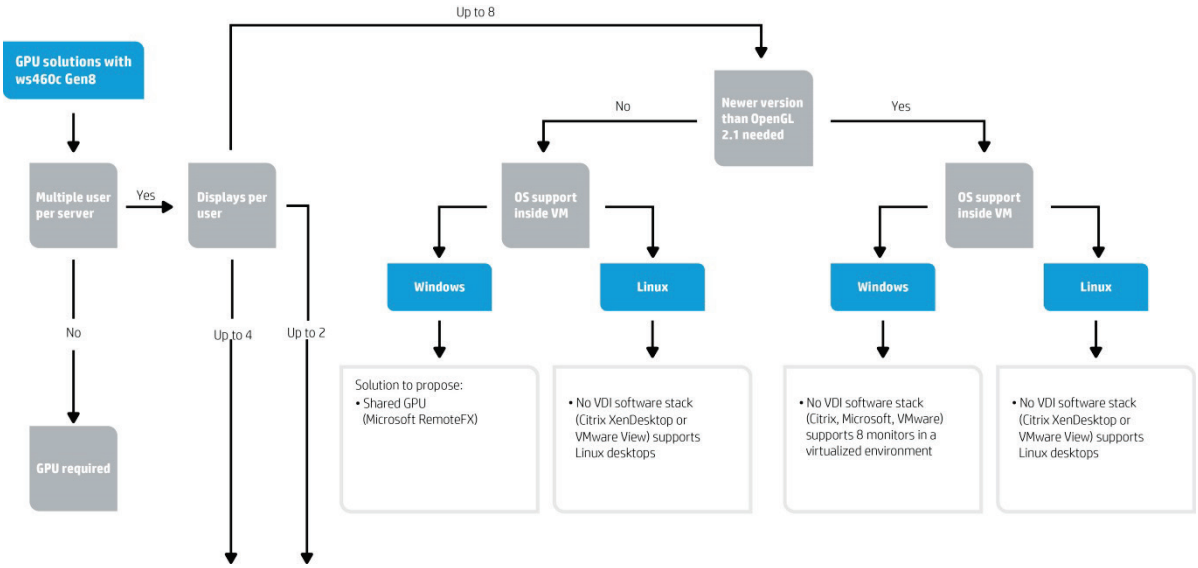
This method uses an API intercept model where the GPU is owned and managed by the hypervisor and all incoming graphics API requests from the VM's are intercepted via the API capture driver in the VM and redirected to and executed by the hypervisor and then sent back to VM (see figure 5). The VM does not have direct access to the GPU, and the GPU driver is loaded within the hypervisor

GPU: NVIDIA Quadro 3000m, 5000, 6000, NVIDIA GRID K1 and K2 Density: depends on workload of each user

Hyper-V RemoteFX considerations

The GPU has a dedicated amount of video RAM, for example, an NVIDIA Quadro 6000 has 6 GB of video RAM. Microsoft RemoteFX (VM with vGPU connected) virtual machines consume a specific amount of video RAM based on the max number of monitors and resolution set for each virtual machine. This will dictate the maximum number of virtual machine per physical GPU. Tables 3 and 4 give you the memory allocation based on monitor and resolution configuration

Figure 16. Decision tree, multiple user per server—Up to 8 displays per user


GPU solutions with ws460c Gen8
Multiple user per server
Yes
No
GPU required
Displays per user
Up to 8
Up to 4
Up to 2
Newer version than OpenGL 2.1 needed
No
Yes
OS support inside VM
OS support inside VM
Windows
Linux
Windows
Linux
Solution to propose:
• Shared GPU (Microsoft RemoteFX)
• No VDI software stack (Citrix XenDesktop or VMware View) supports Linux desktops
• No VDI software stack (Citrix, Microsoft, VMware) supports 8 monitors in a virtualized environment
• No VDI software stack (Citrix XenDesktop or VMware View) supports Linux desktops

Resources

To read more about HP ProLiant WS460c Gen8 Graphics Server Blade, go to hp.com/go/bladeworkstation

To learn more about HP Client Virtualization reference architectures, go to hp.com/go/cv

HP and Citrix, hp.com/go/citrix and citrix.com/hp

Citrix XenDesktop, citrix.com/xendesktop

Citrix XenApp, citrix.com/xenapp

Microsoft Hyper-V, microsoft.com/hyper-v

HP and VMware View, hp.com/go/vmware and vmware.com/view

Learn more at
hp.com/go/cv

Sign up for updates
hp.com/go/getupdated

Share with colleagues

Rate this document

© Copyright 2012–2013 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Intel Xeon and Intel Pentium are trademarks of Intel Corporation in the U.S. and other countries. Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation. ATI is a trademark of Advanced Micro Devices, Inc.

4AA4-1701ENW, August 2013, Rev. 2